BRC4 Meeting
BRCs Program Update

Valentina Di Francesco
Bioinformatics Program Director
Microbial Genomics Program, DMID

BRC4 meeting agenda

Day 1

- "Hot topics" from the 8 centers
- "Other Initiatives"
  - JGI/LANL sequence project
  - Uniprot and Genome Reviews
  - NIAID Administrative Center of the Proteomics Research Centers
  - NIAID Immune Epitope Database
  - NIAID Microbial Sequencing Centers
- Inter-Operability Working Group session

Day 2

- Annotation Working Group session

Presentations will be posted on the NIAID BRC web site (except for those from the AWG)

Logistics

- Breakfast, coffee breaks, lunches and internet access included in the meeting fee.
- On each day @ 10:30am the meeting will move to room 170
- Group Dinner at 7pm tonight at the Thai Surin West
  - Meet in the hotel lobby at 6:45pm

Feb 2006 (BRC3) – Dec 2006 (BRC4)

- All BRCs web sites have been updated
  - Content/Database interfaces/Face lifting
- Annotation SOPs are being developed by all BRCs – First draft due in January 2007
- AWG
  - Established in Summer 2006 - 1st Meeting held in August 2006
  - First round of performance metrics will be presented at BRC4
- IOWG
  - 2 conference calls
  - BRC-Central site last release July 2006
  - GFF3 files, releases of s/w packages and meeting annoucements have been regularly updated
- New BRC additions
  - VBRC
    - ICTVdb http://phene.cpmc.columbia.edu/
    - will start supporting the Hepatitis C Virus
  - BioHealthBase: LANL Influenza Sequence Database
  - PATRIC: SwissProt
- SWGs
  - 4 BRCs have not had a meeting/teleconference since Spring/Summer of 2005

Overview of the GFF3 data
BRC3 vs BRC4

Center	Last Data Upload	Gene Count	Genes With EC Numbers	Genes With GO IDs	Genes With Gene Symbols	Total Number of tRNAs	Total Number rRNAs	Total Number of RNAs	Total Number of GO Ids Assigned To Genes
ApiDB	10/25/2005	12,831	0	0	0	91	39	150	0
	8/18/2006	61,338	2,064	7,127	0	136	62	220	20,022
BHB	10/18/2005	23,003	0	0	0	357	36	401	0
	11/8/2006	23,438	0	8,436	0	348	42	392	20,805
ERIC	1/10/2006	102,106	2,402	0	40,410	1,296	327	1,769	0
	7/6/2006	156,420	8,404	3,176	58,404	2,622	455	3,371	74,556
NMPDR	1/8/2006	63,878	0	0	0	1,929	0	1,929	0
	10/3/2006	123,874	33,054	0	35,479	3,214	0	3,214	72,556
PATRIC	1/26/2006	23,991	0	0	5,447	350	44	408	0
	12/5/2006	31,621	1,683	3,150	4,596	135	19	155	8,611
TIGR	2/1/2006	149,139	28,469	67,085	45,363	0	0	0	137,570
	10/3/2006	159,350	25,175	72,759	34,851	1,669	1,228	2,897	168,318
VBRC	1/18/2006	5,119	0	0	0	0	0	0	0
	11/3/2006	30,001	0	0	0	0	0	0	0
VectorBase	1/11/2006	15,802	0	0	0	0	0	0	0
	8/22/2006	30,428	5,180	18,129	0	1,412	233	1,760	84,793

Challenges

- Demostrate BRCs added value to the scientific community
  - Improvements to the Genbank/Refseq annotations
  - BRCs annotation SOPs
- Collaborations with developers of ‘products’ in the biodefense community
- Outreach

ApiDB workshop - introductory survey of common terminology (June 2006)

Term	Not at all	Heard of it	Slightly familiar	Very familiar
chromosome	0.0	3.8%	15.4%	80.8%
SNPs	11.5%	19.2%	34.6%	30.8%
UTR	23.1%	11.5%	26.9%	38.5%
curated annotation	19.2%	30.8%	30.8%	19.2%
EC numbers	42.3%	19.2%	11.5%	26.9%
GO term	15.4%	26.9%	34.6%	19.2%
PDB structures	46.2%	15.4%	15.4%	19.2%
GenBank	0.0	19.2%	19.2%	57.7%
RefSeq	38.5%	23.1%	34.6%	3.8%%

BRC5 ?

October / November 2007

BRCs Volunteers Wanted in
Suburban Washington DC

Acknowledgments

- Elliot Lefkowitz and University of Alabama at Birmingham
- Owen White (IOWG) and Ross Overbeek (AWG)
- David Bruce (JGI)
- Rolf Apweiler (Uniprot)
- Bjorn Peters (LJAI)
- JoJo Stemple & Peter McGarvey (AC of PRCs)
- Eric Eisenstadt (TIGR MSC) & Matthew Henn (BROAD MSC)

AWG

- Mission
  - To provide **quantitative** measures about the **quantity** and **quality** of the BRCs "added value" that can be monitored **over time**

- By BRC4, each BRC to identify, implement and analyze the metrics
  - create a baseline for measuring progress over time
  - For prokaryotic genomes
    - Measures of consistency, accuracy, completeness of annotations using both FIGFams and TIGRFams
  - Post the metrics on each public BRC site

- Appeal:
  - for the time being, do not share with non-BRCs members comparisons of metrics across BRCs (unless you are prepared to fully and extensively justify the differences among the BRCs).

AWG session presentations will NOT be posted on the NIAID web site

Table 3a. Database Usage / Community Access

(ApiDB-specific)	*CryptoDB*	*PlasmoDB*	*ToxoDB*	*ApiDB*
Database usage				
- bandwidth				
- data downloads				
- users (total, unique)				
- hits (total, unique)				
- queries (total, unique)				
% of sessions distrib'n by query				
- 'history' utilization				
% of sessions				
Outreach				
- publications & CDs				
- lectures/seminars				
- mtg presentations				
- workshops, road-shows, webcasts				

Table 3b. Database Usage / Community Access

(ApiDB-specific)	*CryptoDB*	*PlasmoDB*	*ToxoDB*	*ApiDB*
Feedback **- bug reports** **- feature requests** **- community annot.** **- surveys?**				
Database citations **- Google Scholar** **- publications** **- grants/patents** **- meetings**				
Scale of community	***Cryptosporidium* or cryptosporidiosis**	***Plasmodium* or malaria**	***Toxoplasma* or toxoplasmosis**	**apicomplexa or apicomplexan**
- Google Scholar				
- PubMed / 3 yr				
≥ 5 pubs / 3 yr				
- CRISP / 5 yr				

Questions for today's speakers

- Metrics
  - Results and Discussion?
  - Do you need to modify the metrics or add to them?
  - Are both TIGRFams and FIGFams useful?
    - Was it challenging to use the data generated by both of them?
    - How would you change the FIGFams and TIGRFams data access process?
  - Frequency of updates to metrics?
    - Proposal: quarterly
  - How will you post the metrics on your web site?

- Annotation SOPs
  - How will you post the SOPs on your web site?

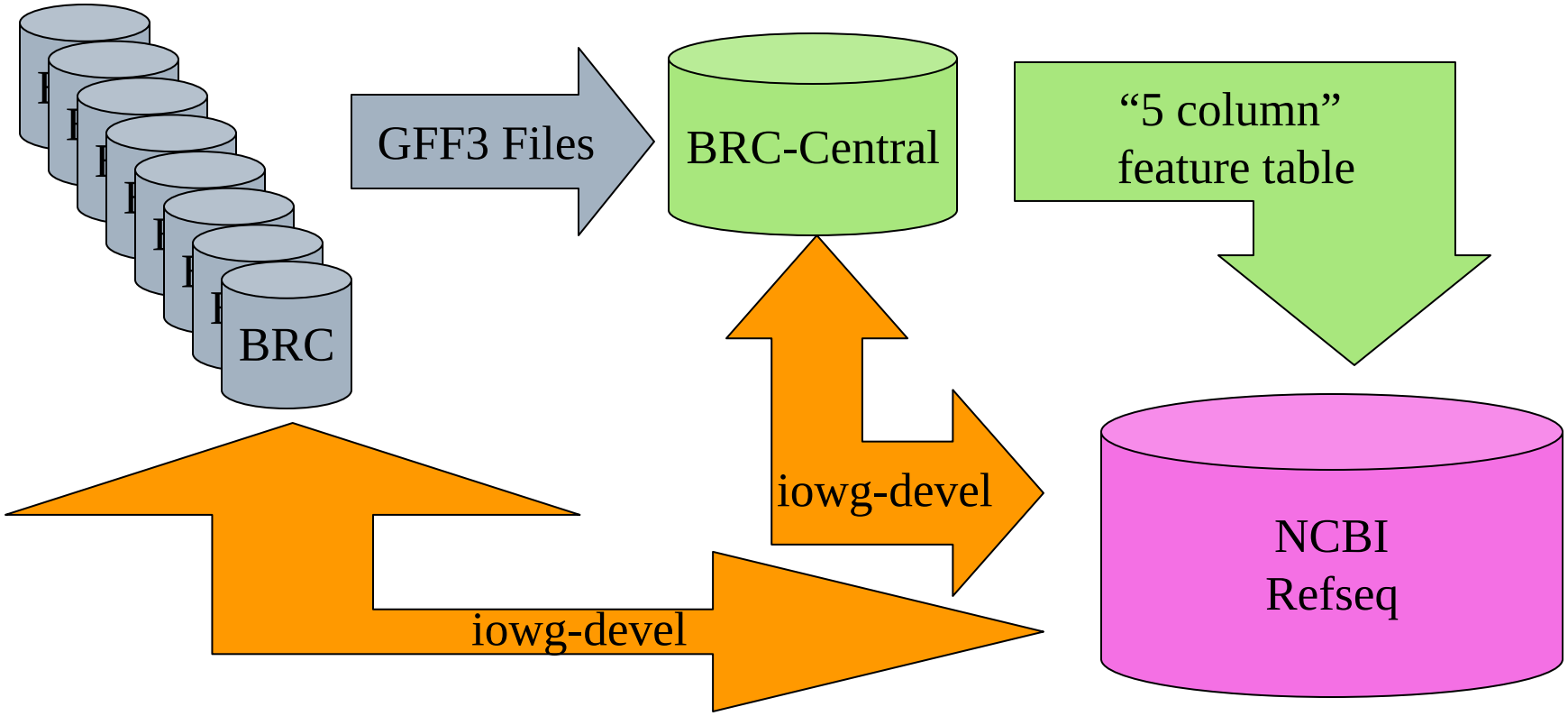
IOWG

- Keeping the community posted on which genome/assembly each BRC is annotating
  - Generate data for a table to be posted on BRC-Central with the following information -
    - Accession numbers and versions of the genomes
    - Primary sequencer
    - Group with the responsibility for updating the Genbank annotation
    - Group reannotating the genome
      - gene boundaries/merges/deletions
      - Adding genome features (i.e. pseudogenes, RNAs, repeats, IS elements, etc.)
      - Functional annotation (GO, E.C., Gene names, subsystems or specific protein families)
    - Which SOPs are available
  - Table to be posted on BRC-Central by January 31, 2007

BRCs submissions to **Refseq**

- Why?
  - increase visibility of BRCs through link-outs from NCBI
  - The users need it – it reduces confusion
- Background information
  - Refseq has been anxiously waiting for the BRCs to submit genome annotation updates that would then become Refseq records
  - Refseq is looking forward to receiving other types of annotations (i.e. PATRIC's PIML files, rich annotations of the WNV genes)
  - Several attempts have been made by Refseq reps to request data from the BRCs with little success
  - Hence NCBI has downloaded the entire set of GFF3 files from BRC-Central, but could not use the information:
    - Missing: reference to a source genome accession and version number
    - Errors in the files
    - Inconsistencies in the use of the tags
- Caveat: NCBI reserves the right to reject some annotation updates from the BRCs

Data and Communications Flow



Some considerations and action items

- Will use a process that is already in place at each BRC to generate GFF3 files: the additional burden is on BRC-Central to generate the tables for NCBI
- Need to “enrich” the information contained in the GFF3 files
  - Need to consistently use accession.version of the genomes and genes
  - Need to have robust QA checks of the files content at BRCs and BRC-Central (i.e. checks on gene coordinates)
  - Need to flag modified genes?
- First set of NCBI-Format files ready by January 31, 2007
- Centralized communication between BRC-Central and NCBI